A Novel Method for Speaker Independent Recognition Based on Hidden Markov Model

Feng-Long Huang
Computer Science and Information Engineering, National United University
No. 1, Lienda, Miaoli, Taiwan, 36003
flhuang@nuu.edu.tw

Abstract: **In this paper, we address the speaker independent recognition of Chinese number speeches 0~9 based on HMM. Our former results of inside and outside testing achieved 92.5% and 76.79% respectively. To improve further the performance, two important features of speech; MFCC and cluster number of vector quantification, are unified together and evaluated on various values. The best performance achieve 96.2% and 83.1% on MFCC Number = 20 and VQ clustering number = 64.**

Keywords: **Speech Recognition, Hidden Markov Model, LBG Algorithm, Mel-frequency cepstral coefficients, Viterbi Algorithm.**

I. Introduction

In Speech processing, automatic speech recognition (ASR) is capable automatically of understanding the input of human speech for the text output with various vocabularies. ASR can be applied in a wide range of applications, such as: human interface design, speech Information Retrieval (SIR) [11,12], language translation, and so on. In real world, there are several commercial ASR systems, for example, IBM's Via Voice, Mandarin Dictation System–the Golden Mandarin (III) of NTU in Taiwan, Voice Portal on Internet and 104 on-line speech queries systems. Modern ASR technologies merged the signal process, pattern recognition, network and telecommunication into a unified framework. Such architecture can be expanded into broad domains of services, such as e-commerce and wireless speech system of WiMAX.

The approaches adopted on ASR can be categorized as: 1)Hidden Markov Model (HMM) [1,2,3,4], 2)Neural Networks [5,6,7], 3)Wavelet-based and spectrum coefficients of speech [15,16], other method is the combination of first two approaches above [8,9]. The Hidden Markov Model is a result of the attempt to model the speech generation statistically, and thus belongs to the first category above. During the past several years it has become the most successful speech model used in ASR. The main reason for this success is the powerful ability to characterize the speech signal in a mathematically tractable way.

In a typical ASR system based on HMM, the HMM stage is proceeded by the parameter extraction. Thus the input to the HMM is a discrete time sequence of parameter vectors, which will be supplied to the HMM.

In the paper, the following sections are organized as follow: the process of speeches is introduced in Section 2 and the acoustic model of recognition will be described in Section 3. The initial results for former approaches are presented in Section 4. The improvement metods are furthermore described in Section 5

II. Processes of Speech

In this section, we will describe all the procedures for pre-processes.

A. Processing Speech

The analog voice signals are recorded thru microphone. It should be digitalized and quantified. The digital signal process can be described as follows:

$$x_p(t) = x_a(t)\,p(t) \quad (1)$$

where $x_p(t)$ and $x_a(t)$ denote the processed and analog signal. $p(t)$ is the impulse signal.

Each signal should be segmented into several short frames of speech which contain a time series signal. The features of each frame are extracted for further processes.

B. Pre-emphasis

Basically, the purpose of pre-emphasis is to increase, the magnitude of some (usually higher) frequencies with respect to the magnitude of other (usually lower) frequencies in order to improve the overall signal-to-noise ratio (SNR) by minimizing the adverse effects of such phenomena as attenuation distortion.

C. Frame Blocking

While analyzing audio signals, we usually adopt the method of short-term analysis because most audio signals are relatively stable within a short period of time. Usually,

DOI: 01.IJSIP.02.01.218

the signal will be segmented into time frame, say 15 ~ 30 ms.

D. Hamming Window

In signal processing, the window function is a function that is zero-valued outside of some chosen interval. The Hamming window is a weighted moving average transformation used to smooth the periodogram values.

Supposed that original signal $s(n)$ is as follows:

$$s(n),\ n = 0,\ldots N\text{-}1 \quad (2)$$

The original signal $s(n)$ is multiplied by hamming window $w(n)$, we will obtain $s(n)* w(n)$, $w(n)$ can be defined as follows:

$$w(n) = (1 - \alpha) - \alpha*\cos(2\pi n/(N\text{-}1)),\ 0 \leqq n \leqq N\text{-}1 \quad (3)$$

where N denotes the sample number in a window.

E. Mel-frequency cepstral coefficients

Mel Frequency Cepstral Coefficient (MFCC) is one of the most effective feature parameter in speech recognition. For speech representation, **i**t is well known that MFCC parameters appear to be more effective than power spectrum based features. MFCCs are based on the human ears' non-linear frequency characteristic and perform a high recognition rate in practical application.

- lower frequency, human hear more acute.
- higher frequency, human hear less acute.

As shown in Fig. 7, MFCC are presented as:

$$mel(f)=1125*\ln(1+f/700) \quad (4)$$

III. Acoustic Model of Recognition

A. Vector Quantification

Foundational vector quantifications (VQ) were proposed by Y. Linde, A. Buzo, and R. Gray in 1980, So-called LBG algorithm. LBG is based on k-means clustering [2,5], referring to the size of codebook G, training vectors will be categorized into G groups. The centroid Ci of each Gi will be the representative for such vector of codeword. In principal, the category is tree based structure.

B. Hidden Markov Model

A Hidden Markov Model (HMM) is a statistical model in which is assumed to be a Markov process with unknown parameters. The challenge is to find all the appropriate hidden parameters from the observable states. HMM can be considered as the simplest dynamic Bayesian network.

In a regular Markov model, the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. However, in a hidden Markov model, the state is not directly visible (so-called *hidden*), while the variables influenced by the state are visible. Each state has a probability distribution over the output. Therefore, the sequence of tokens generated by an HMM gives some information about the sequence of states.

A complete HMM can be defined as follows:

$$\lambda = (\pi, A, B) \quad (5)$$

HMM model can be defined as (π , A, B) :

1. **Π** (Initial state probability):

$$\pi = \{\pi_i = \text{prob}(q_1 = S_i)\} \qquad 1 \leq i \leq N \quad (6)$$

2. **A (**State transition probability):

$$A = \{a_{ij} = \text{prob}(q_{t+1} = S_j \mid q_t = S_i)\} \quad (7)$$
$$1 \leq i \leq N$$

3. **B (**Observation symbol probability):

$$B = \{b_j(O_t) = \text{prob}(O_t \mid q_t = S_j)\} \quad 1 \leq i \leq N \quad (8)$$

where $O = \{O_1, O_2, \ldots, O_T\}$ is the observation.

$S = \{S_1, S_2, S_3, \ldots, S_N\}$ is state symbols and

$q = \{q_1, q_2, q_3, \ldots, q_T\}$ is observation states and T denote the length of observation, N is the number of states.

C. System Models

The recognition system is composed of two main functions: 1) extracting the speech features, including frame blocking, VQ, and so on, 2) constructing the model and recognition based on the HMM, VQ and Viterbi Algorithm.

It is apparent that short speech signal varied sharply and rapidly, whereas longer signal varied slowly. Therefore, we use the dynamic frame blocking rather than fixed frame for different experiments.

IV. Initial Experiments

A. Recognition System Based on HMM

In the paper, we focus on speaker independent speech recognition of Chinese number speeches 0~9. All the samples with 44100 Hz/16 bits are recorded by three native male adults. Total 560 samples are divided into two parts, 280 for training and 280 for testing. After complete the pre-process, such as preemphasis, frame boloking, VQ.

B. Comparison for fixed and Dynamic Frame Size

According to our empirical results, comparing the fixed and dynamic frame size, recognition rate of fixed

DOI: 01.IJSIP.02.01.218

frame size achieves 76.79%, and superior to the other with75.71%, as shown in Table 1.

Table 1: comparing the frame size, **(SymbolNum=64)**

		wave Num	Mfcc time	VQ time	HMM training	Symbol Num	rate(%)
fixed	I	280	32.9	5.77	3.44	64	90.36
	O	280					**76.79***
dynamic	I	280	32.0	3.31	2.42	64	92.50*
	O	280					75.71

PS. I and O denote the inside and outside testing, respectively

V. Further Improvement

A. Improving the Samples of Speech

According to our empirical results, recognition rate achieve better results while cluster number=64. Inside and outside testing are 92.5% and 76.79%, respectively.

To improve the performance, we analyze all the speech wavelet. There are many samples affected by boost noise derived from human speaking or environment, as shown in Fig. 1. In such a situation, the end points of boosted speech cannot be usually detected correctly. It will lead to degrade the performance of system.

Usually, detecting end points judged on ZCR and energy of speech, as shown in Fig. 1. However, it is significant that we need extra features to detect for noise situation. Based on experimental results and observation, the improvement rules are summarized as follows:

Input: X(n) , n = 1 to j
Output: Y(m),1 <= m <= j

1. segment the speech X(n): framedY = framed (X(n))
2. calculate the ZCR and energy for each frame.
3. smooth the curves for both ZCR and energy
4. calculate the average of first 10 frames, and multiplying 1.2. The average value will be used as the threshold for detecting process.
5. ZCR is valid only if framedY is larger than 100, as shown in Fig. 2.
6. the speech will be effective only if the size is larger than 3ms.
7. the starting energy of speech should be larger than threshold.
8. the energy for continuous 5 frames of speech should be increased progressively.

Referring to the improvement, the speeches number 8 (ㄅㄚ) with boost noise can be detected, as shown in Fig. 2. The improvement of detection will leads to better results for following recognition process.

B. Better Combination of Various Features

To improve furthermore the performance, two spectrum features, MFCC and cluster number, of speeches are unified and evaluated. MFCC degree varied from 8 to 36 with interval 4 and cluster number varied on 32 to 256 with interval 32. We evaluated all the combination for these two features with various numbers. The process times needed for computation are shown in Table 2. The best results can achieve on MFCC Number= 20 and VQ clustering number = 64. The inside and outside testing of recognition achieve 96.2% and 83.1% shown in Fig. 3 and net results for inside and outside testing are 3.7% and 6.3% respectively. We just list the results with VQ = 64 in the paper.

Table 2: processed time with VQ = 64.

MFCC degree	**8**	**12**	**16**	**20**	**24**	**28**	**32**	**36**
MFCC	15.8	16.9	18.6	23.5	25.3	27.2	28.5	29.9
VQ	1.0	2.6	3.3	3.4	3.8	4.9	5.3	6.6
HMM	1.7	1.7	1.8	1.8	1.8	1.8	1.9	1.9

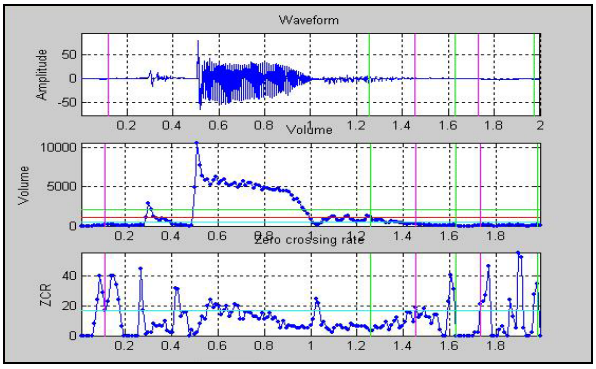

Fig. 1: before improvement, Chinese number 8 (ㄅㄚ)

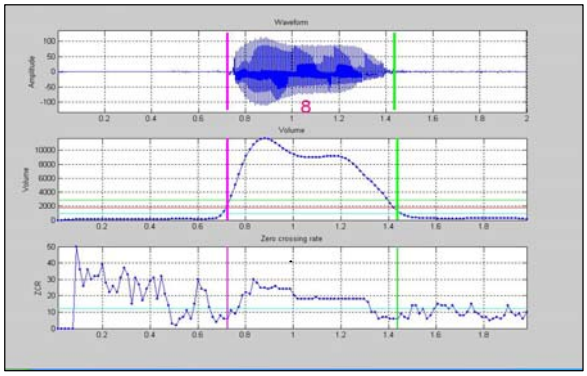

Fig. 2: after improvement, Chinese number 8 (ㄅㄚ).

DOI: 01.IJSIP.02.01.218

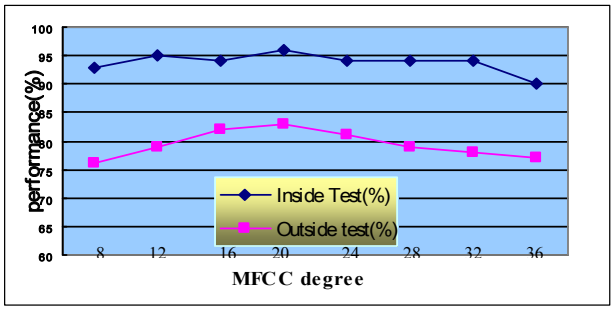

Fig. 3: performance with VQ = 64, MFCC degrees varied between 8 and 36.

VI. Conclusion

In this paper, we address the speaker independent speech recognition of Chinese number speeches based on HMM. The algorithm for our novel approach is proposed for the speech recognition. 480 speech samples are recorded and pre-processed. The preliminary results of outside testing achieve 76.79%.

To improve furthermore the performance, two features of speeches; MFCC and VQ cluster number, are evaluated. We then find the combination of two spectrum features to achieve best results. The best performance will be achieved on MFCC, Number = 20 and VQ clustering number = 64. The final inside and outside testing of recognition achieve 96.2% and 83.1%. It proves that the proposed approach can be employed to recognize the speaker independent speeches.

Future works will be studied in the following:

1) Employing other effective methods to merging novel method to enhance the performance.
2) Applying the method into isolated Chinese speech recognition.
3) Improving the precision rates.

Acknowledgement

The paper is supported under the Project of Lein-Ho Foundation, Taiwan.

References

[1] Keng-Yu Lin, 2006, Extended Discrete Hidden Markov Model and Its Application to Chinese Syllable Recognition, Master thesis of NCHU, Taiwan.

[2] Keng-Yu Lin, 2006, Extended Discrete Hidden Markov Model and Its Application to Chinese Syllable Recognition, Master thesis of NCHU.

[3] X. Li, M. Parizeau and R. Plamondon, April 2000, Training Hidden Markov Models with Multiple Observations--A Combinatorial Method, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol. 22, No. 4.

[4] A. Sperduti and A. Starita, May 1997, Supervised Neural Networks for Classification of Structures. IEEE Transactions on Neural Networks, 8(3): pp.714-735.

[6] E. Behrman, L. Nash, J. Steck, V. Chandrashekar, and S. Skinner, October 2000, Simulations of Quantum Neural Networks, Information Sciences, 128(3-4): pp. 257-269.

[7] Hsien-Leing Tsai, 2004, Automatic Construction Algorithms for Supervised Neural Networks and Applications, PhD thesis of NSYSU, Taiwan.

[8] Li-Yi Lu, 2003, The Research of Neural Network and Hidden Markov Model Applied on Personal Digital Assistant, Master thesis of CYU, Taiwan.

[10] Rabiner, L. R., 1989, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proceedings of the IEEE, Vol.77, No.22, pp.257-286.

[11] Manfred R. Schroeder, H. Quast, H.W. Strube, Computer Speech: Recognition, Compression, Synthesis , Springer, 2004.

[12] Wald, M., 2006, Learning Through Multimedia: Automatic Speech Recognition Enabling Accessibility and Interaction. Proceedings *of ED-MEDIA 2006:* World Conference on Educational Multimedia, Hypermedia & Telecommunications. pp. 2965-2976.

[13]A. Revathi, R. Ganapathy and Y. Venkataramani, Nov. 2009, Text Independent Speaker Recognition and Speaker Independent Speech Recognition Using Iterative Clustering Approach, International Journal of Computer science & Information Technology (IJCSIT), Vol. 1, No 2, pp.30-42.

[14]Haamid M. Gazi, Omar Farooq, Yusuf U. Khan, Sekharjit Datta, 2008, Wavelet-based, speaker-independent isolated Hindi digit recognition International Journal of Information and Communication Technology, Vol. 1 , Issue 2 pp. 185-198

[15]Chakraborty P., et at., 2008, An Automatic Speaker Recognition System, Neural Information Processing, Lecture Notes in Computer Science (LNCS), Springer Berlin / Heidelberg, pp. 517-526.

[16] Kun-Ching Wang, 2009, Wavelet-Based Speech Enhancement Using Time-Frequency Adaptation, EURASIP Journal on Advances in Signal Processing, Volume 2009 (2009), Article ID 924135.

DOI: 01.IJSIP.02.01.218